Chapter 18
Cosine Similarity-Based Classifiers for Functional Data

Tianming Zhu and Jin-Ting Zhang

Abstract In many situations, functional observations in a class are also similar in shape. A variety of functional dissimilarity measures have been widely used in many pattern recognition applications. However, they do not take the shape similarity of functional data into account. Cosine similarity is a measure that assesses how related are two patterns by looking at the angle instead of magnitude. Thus, we generalize the concept of cosine similarity between two random vectors to the functional setting. Some of the main characteristics of the functional cosine similarity are shown. Based on it, we define a new semi-distance for functional data, namely, functional cosine distance. Combining it with the centroid and k-nearest neighbors (kNN) classifiers, we propose two cosine similarity-based classifiers. Some theoretical properties of the cosine similarity-based centroid classifier are also studied. The performance of the cosine similarity-based classifiers is compared with some existing centroid and kNN classifiers based on other dissimilarity measures. It turns out that the proposed classifiers for functional data perform well in our simulation study and a real-life data example.

18.1 Introduction

Functional data consists of functions. In recent decades, it is prevalent in many fields such as economics, biology, finance, and meteorology (for an overview, see [14]). The goals of the functional data analysis (FDA) are essentially the same as those of any other branch of statistics [13]. References [5, 13] provided broad overviews of the techniques of FDA. In this paper, we are interested in supervised classification for functional data.

T. Zhu · J.-T. Zhang (✉)
Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore
e-mail: stazjt@nus.edu.sg

T. Zhu
e-mail: stazt@nus.edu.sg

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_18

Supervised classification is one of the oldest statistical problems in experimental science. We have a training sample and a test sample whose class memberships are known. The aim of classification is to create a method for assigning a new coming observation to one of the predefined classes based on the training sample. Its classification accuracy can be assessed via the misclassification error rate (MER) of the test sample. Many supervised classification methods for functional data have been developed in recent years. A number of studies have extended the traditional classification methods for multivariate data to the context of functional data. For instance, [1] proposed to filter the training samples of functional observations using the Fourier basis so that the classical kNN classifier can be applied to the resulting Fourier coefficients. [15] extended the methodology based on support vector machine for functional data. In addition, a centroid method for classifying functional observations has been adopted by [3]. They used the project of each functional observation onto a given direction instead of the functional observation itself so that a functional data classification problem becomes a one-dimensional classification problem. Further, [11] extended linear discriminant analysis to functional data. References [8–10, 17] proposed classifiers based on functional principal components while [4] developed functional classifiers based on shape descriptors.

The concepts of similarity and distance are fundamentally important in almost every scientific field. Similarity and distance measures are also an essential requirement in almost all pattern recognition applications including classification, clustering, outlier detection, regression and so on. There exist a large number of similarity measures in the literature and the performance of any pattern recognition technique largely depends on the choice of the similarity measures. In the recent literature on functional data, some authors have proposed semi-distances well adapted for sample functions such as the semi-distances based on functional principal components [5] and the functional Mahalanobis semi-distance [7]. However, most of the similarity measures are used in multivariate data and have not been extended to the functional framework. Our first contribution is to extend the cosine similarity to functional settings and define a new semi-distance for functional data.

The cosine similarity measure can be defined between two functional observations. If these two functional observations are similar in shape, this functional cosine similarity measure will be close to 1; if they are not similar or even opposite in shape, the associated cosine similarity measure will be small or even be negative. Therefore, it can be used to classify functional data. Our second contribution is that by combining the new functional semi-distance with the centroid and kNN classifiers, we propose the cosine similarity-based classifiers for functional data.

The rest of this work is organized as follows. We review a number of dissimilarity measures for functional data in Sect. 18.2. Section 18.3 introduces the concept of functional cosine similarity (FCS) and shows its main characteristics. Based on FCS, we define functional cosine distance (FCD). Section 18.4 develops the FCD-based centroid and kNN classifiers for functional data. In particular, the asymptotic MER of the FCD-based centroid classifier for functional data is derived. A simulation study for comparing the proposed cosine similarity-based centroid and kNN classifiers against other existing centroid and kNN classifiers is presented in Sect. 18.5.

Applications of the proposed cosine similarity-based centroid and kNN classifiers to a real-life data example is given in Sect. 18.6. Some concluding remarks are given in Sect. 18.7. The proofs of the main theoretical results are given in the Appendix.

18.2 Functional Dissimilarity Measures

In this section, we review some dissimilarity measures for functional data. In practice, functional data are obtained via observing some measure over time, and we assume the sample of functional observations was generated from a stochastic process.

Let $\mathscr{T}$ be some compact set. Let $x(t), t \in \mathscr{T}$ be a stochastic process having mean function $\eta(t), t \in \mathscr{T}$ and covariance function $\gamma(s,t), s,t \in \mathscr{T}$. We write $x(t) \sim \mathrm{SP}(\eta, \gamma)$ for simplicity. Throughout this work, let $\mathscr{T}$ be a finite interval, and we use $\|x\|_p$ to denote the L^p-norm of a function $x(t), t \in \mathscr{T}$: $\|x\|_p = \left(\int_{\mathscr{T}} |x(t)|^p dt\right)^{1/p}$, for $p = 1, 2, \ldots$. When $p = 2$, we may use $\|\cdot\|$ to denote the L^2-norm for simplicity. If $\|x\|_p < \infty$, we say $x(t), t \in \mathscr{T}$ is L^p-integrable. In this case, we write $x \in \mathscr{L}^p(\mathscr{T})$ where $\mathscr{L}^p(\mathscr{T})$ denotes the Hilbert space formed by all the L^p integrable functions over $\mathscr{T}$. In particular, $\mathscr{L}^2(\mathscr{T})$ denotes the Hilbert space formed by all the squared integrable functions over $\mathscr{T}$, which is an inner product space. The associated inner-product for any two functions in $\mathscr{L}^2(\mathscr{T})$ is defined as $< x, y >= \int_{\mathscr{T}} x(t)y(t)dt,\ x(t), y(t) \in \mathscr{L}^2(\mathscr{T})$. The above L^p-norm and inner-product definitions can be used to define various dissimilarity measures. Let $x(t)$ and $y(t)$ be two functional observations defined over $\mathscr{T}$, which are L^p integrable. The L^p-distance between $x(t)$ and $y(t)$ is then defined as:

$$d_p(x, y) = \|x - y\|_p,$$

for $p = 1, 2, \ldots$. We often use L^1, L^2, and L^∞-distances. It is well known that $d_\infty(x, y) = \|x - y\|_\infty = \sup_{t \in \mathscr{T}} |x(t) - y(t)|$.

The L^p-distances can be implemented easily in supervised classification but they do not take the correlation of a functional observation into account. To partially address this issue, [7] proposed the so-called functional Mahalanobis semi-distance so that the correlation structure of functional observations can be taken into account partially. The functional Mahalanobis semi-distance is defined using a number of the largest eigenvalues and the associated eigenfunctions. Note that when the covariance function $\gamma(s,t)$ has a finite trace, i.e., $\mathrm{tr}(\gamma) = \int_{\mathscr{T}} \gamma(t,t)dt < \infty$, it has the following singular value decomposition ([18], p. 3): $\gamma(s,t) = \sum_{r=1}^{\infty} \lambda_r \phi_r(s)\phi_r(t)$, where $\lambda_r, r = 1, 2, \ldots$ are the decreasing-ordered eigenvalues of $\gamma(s,t)$, and $\phi_r(t)$, $r = 1, 2, \ldots$ are the associated orthonormal eigenfunctions.

Let $y(t) \sim \mathrm{SP}(\eta, \gamma)$. By assuming $\gamma(s,t)$ has a finite trace, we have the following Karhunen-Loéve expansion: $y(t) = \sum_{r=1}^{\infty} \xi_r \phi_r(t)$, where $\xi_r =< y, \phi_r >, r = 1, 2, \ldots$ denote the associated principal component scores of $y(t)$. Let $x(t)$ be another functional observation whose covariance function is also $\gamma(s,t)$. Then we can also

expand $x(t)$ in terms of the eigenfunctions of $\gamma(s,t)$ as $x(t) = \sum_{r=1}^{\infty} \zeta_r \phi_r(t)$, where $\zeta_r =< x, \phi_r >, r = 1, 2, \ldots$ denote the associated principal component scores of $x(t)$. Then, the functional Mahalanobis (FM) semi-distance between $x(t)$ and $y(t)$ is given by

$$d_{FM,q}(x, y) = \left(\sum_{r=1}^{q} \lambda_r^{-1} (\zeta_r - \xi_r)^2 \right)^{1/2}.$$

Based on the principal component scores of $x(t)$ and $y(t)$, [5] defined the so-called functional principal components (FPC) based semi-distance which can be used as a dissimilarity measurement:

$$d_{FPC,q}(x, y) = \left(\sum_{r=1}^{q} (\zeta_r - \xi_r)^2 \right)^{1/2}.$$

Based on these dissimilarity measures, a number of classifiers are adopted for functional data. However, all these dissimilarity measures do not take the shape similarity of the functional data into account. Note that in many situations, functional observations in one class are also similar in shape. To take this information into account, in the next section, we introduce the cosine similarity measure for functional data.

18.3 Functional Cosine Similarity

The main goal of this section is to generalize the cosine similarity measure between two random vectors to the functional settings. The cosine similarity measure between two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as: $\mathrm{CS}(\mathbf{x}, \mathbf{y}) = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\|\|\mathbf{y}\|}$, where $\|\cdot\|$ and $<\cdot,\cdot>$ denote the usual Euclidean norm and the usual inner product in $\mathbf{R}^n$. It is seen that the cosine similarity measure is the ratio of the inner product between the two vectors to the product of their Euclidean norms. The main characteristic of the cosine similarity measure is that it measures the closeness or similarity between two vectors using the cosine value of the angle between the two vectors, which takes value between $[-1, 1]$. It is thus a judgment of orientation and not magnitude. If two vectors have the same orientation, they have a cosine similarity measure of 1; if two vectors are orthogonal, they have a cosine similarity measure of 0; if two vectors have exactly opposite orientations, they have a cosine similarity measure of -1. When two vectors are similar, this similarity measure will take larger values.

We now extend the above cosine similarity measure to for functional data. Let $x(t), t \in \mathscr{T}$ and $y(t), t \in \mathscr{T}$ be any two functions in $\mathscr{L}^2(\mathscr{T})$. Then the functional cosine similarity (FCS) measure between $x(t)$ and $y(t)$ can be defined as follows:

$$\mathrm{FCS}(x, y) = \frac{< x, y >}{\|x\|\|y\|},$$

where $\|\cdot\|$ and $<\cdot,\cdot>$ denote the usual L^2-norm and the usual inner product in $\mathscr{L}^2(\mathscr{T})$ as defined before. It is seen that $\mathrm{FCS}(x, y)$ measures the similarity or closeness between $x(t)$ and $y(t)$ using the cosine value of the angle between the two functions $x(t)$ and $y(t)$ which was proposed by [13]. It has the following properties: (1) $-1 \leq \mathrm{FCS}(x, y) \leq 1$, normalization; (2) $\mathrm{FCS}(x, y) = \mathrm{FCS}(y, x)$, symmetry or commutativity; (3) $x(t) = y(t) \Rightarrow \mathrm{FCS}(x, y) = 1$, reflexivity; and (4) $\mathrm{FCS}(x, y) =< \tilde{x}, \tilde{y} >= 1 - \|\tilde{x} - \tilde{y}\|^2/2$ where $\tilde{x}(t) = x(t)/\|x\|$ denotes the normalization version of $x(t)$ and $\tilde{y}(t)$ is similarly defined.

Item (1) says that $\mathrm{FCS}(x, y)$ ranges from -1 (when $x(t)$ is exactly opposite to $y(t)$) to 1 (when $x(t)$ and $y(t)$ are proportional, that is, when $x(t) = ay(t)$) and takes value 0 when $x(t)$ and $y(t)$ are orthogonal. It is due to the fact that $-\|x\|\|y\| \leq < x, y > \leq \|x\|\|y\|$ by the well-known Cauchy-Schwarz inequality between two squared-integrable functions. Items (2) and (3) are obviously held. Item (4) can be shown via some simple algebra. It says that the cosine similarity measure between $x(t)$ and $y(t)$ is exactly 1 minus half of the squared L^2-norm of the difference between their normalization versions $\tilde{x}(t)$ and $\tilde{y}(t)$. Note that $\tilde{x}(t)$ is also called the spatial sign function of $x(t)$ [16], which can be interpreted as the direction of $x(t)$. Thus, the functional cosine similarity measure $\mathrm{FCS}(x, y)$ also can be interpreted as the similarity measure between the directions of $x(t)$ and $y(t)$. If $\tilde{x}(t) = \tilde{y}(t)$, that is, $x(t)$ and $y(t)$ have the same direction, the associated $\mathrm{FCS}(x, y)$ takes value 1.

Note that FCS is not a distance or semi-distance since it is not nonnegative and its value is not 0 when the two functions $x(t)$ and $y(t)$ are exactly the same. However, this can be easily corrected. For this purpose, we define the following functional cosine distance (FCD) between two functions $x(t), t \in \mathscr{T}$ and $y(t), t \in \mathscr{T}$:

$$\mathrm{FCD}(x, y) = [2 - 2\mathrm{FCS}(x, y)]^{1/2} = \left(2 - 2\frac{< x, y >}{\|x\|\|y\|}\right)^{1/2} = \|\tilde{x} - \tilde{y}\|. \quad (18.1)$$

It is obvious that $\mathrm{FCD}(x, y) = 0$ if $x(t)$ and $y(t)$ are exactly the same. Further, we have (1) $0 \leq \mathrm{FCD}(x, y) \leq 2$; (2) $\mathrm{FCD}(x, y) = \mathrm{FCD}(y, x)$, symmetry; and (3) $\mathrm{FCD}(x, y) \leq \mathrm{FCD}(x, z) + \mathrm{FCD}(y, z)$ for any three functions $x(t), t \in \mathscr{T}$, $y(t), t \in \mathscr{T}$ and $z(t), t \in \mathscr{T}$, triangle inequality.

Using the properties of $\mathrm{FCS}(x, y)$, it is easy to verify the first two properties of $\mathrm{FCD}(x, y)$ above. Item (3) can be shown by the well-known Minkowski inequality. Consequently, FCD is a functional semi-distance since $\mathrm{FCD}(x, y) = 0$ cannot imply $x(t) = y(t), t \in \mathscr{T}$. Nevertheless, we can define some classifiers based on FCD for functional data.

18.4 Cosine Similarity-Based Classifiers for Functional Data

Let $G \geq 2$ be an integer. Suppose we have G training functional samples

$$x_{i1}(t), x_{i2}(t), \ldots, x_{in_i}(t) \overset{\text{i.i.d.}}{\sim} \text{SP}(\eta_i, \gamma_i), \quad i = 1, \ldots, G, \tag{18.2}$$

where $\eta_i(t)$'s are the unknown group mean functions and $\gamma_i(s, t)$'s are the unknown group covariance functions. Note that throughout this work, we assume that the functional observations of the same group are i.i.d. and functional observations of different groups are also independent. For a new coming functional observation $x(t)$, our aim is to determine the class membership of $x(t)$ based on the above G training samples.

In this section, our aim is to propose new nonparametric classifiers via combining the centroid and kNN classifiers with FCD. The resulting classifiers are called the FCD-based centroid and kNN classifiers, respectively.

18.4.1 FCD-Based Centroid Classifier

There are many different approaches which can design a nonparametric classifier. The first one, also the simplest one, is based on the concept of similarity. Observations that are similar should be assigned to the same class. Thus, once the similarity measure is established, the new coming observation can be classified accordingly. The choice of the similarity measure is crucial to the success of this approach. The first representative of this approach is the nearest mean classifier, also called nearest centroid classifier. Each class is represented by its mean of all the training patterns in that class. A new observation will be assigned to the class whose mean is closest to the new observation.

For functional data, the class center is the group mean function which can be estimated using its usual group sample mean function. For the G training functional samples (18.2), the G class centers can be estimated as $\bar{x}_i(t) = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}(t)$, $i = 1, \ldots, G$. Then the FCDs between the new coming functional observation $x(t)$ and the above class centers $\bar{x}_i(t), i = 1, \ldots, G$ can be expressed as $\text{FCD}(x, \bar{x}_i), i = 1, \ldots, G$. The FCD-based centroid classifier for functional data then puts $x(t)$ into Class g where

$$g = \text{argmin}_{1 \leq i \leq G} \text{FCD}^2(x, \bar{x}_i). \tag{18.3}$$

18.4.2 FCD-Based kNN Classifier

The classical kNN classifier was first proposed by [6]. Due to its simplicity and efficiency, it is widely used to perform supervised classification in multivariate settings. The classical kNN classifier consists of the following steps: given a training sample with known class labels, classify a new observation into a class by examining its k nearest neighbors and applying the majority vote rule.

For the G training functional samples (18.2), the FCDs between the coming functional observation $x(t)$ and all the training functional observations can be computed as $\mathrm{FCD}(x, x_{ij}),\ j = 1, \ldots, n_i;\ i = 1, \ldots, G$. Let k be some given integer. The $x_{ij}(t)$'s associated with the k smallest values of the above FCDs are the k nearest neighbors of $x(t)$ from the whole training functional sample. Let m_i denote the number of the nearest neighbors from Class i where $i = 1, \ldots, G$. Then $\sum_{i=1}^{G} m_i = k$. Note that some of m_i's are equal to each other and some are equal to 0. The FCD-based kNN classifier for functional data then puts $x(t)$ into Class g where $g = \mathrm{argmax}_{1 \le i \le G} m_i$.

18.4.3 Theoretical Properties of the FCD-Based Centroid Classifier

In this subsection, we study the theoretical properties of the FCD-based centroid classifier. That is, we shall derive its asymptotic misclassification error rate (MER) and show some of its good properties. Recall that $x(t)$ denotes the new coming functional observation. As mentioned in the previous subsection, for the G training functional samples (18.2), the FCD-based centroid classifier will put $x(t)$ to Class g determined by (18.3). For each class $g = 1, \ldots, G$, we have a classification vector function $\mathbf{T}_g(x)$ based on the G-class FCD-based centroid classifier which can be expressed as $\mathbf{T}_g(x) = [T_{g,1}(x), \ldots, T_{g,g-1}(x), T_{g,g+1}(x), \ldots, T_{g,G}(x)]^T$, where $T_{g,i}(x) = \mathrm{FCD}^2(x, \bar{x}_i) - \mathrm{FCD}^2(x, \bar{x}_g)$ for $i = 1, \ldots g-1, g+1, \ldots, G$. Then the G-class FCD-based centroid classifier for functional data assigns $x(t)$ to class g if $\mathbf{T}_g(x) > \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector.

Let π_i denote the probability that $x(t)$ from Class i for $i = 1, \ldots, G$. Assuming that $\mathrm{tr}(\gamma_i) < \infty$, $i = 1, 2, \ldots, G$, we can show that as $n_i, i = 1, 2, \ldots, G$ tend to infinity with $n_i/n \to \tau_i > 0$ where $n = n_1 + n_2 + \cdots + n_G$, we have $\bar{x}_i(t) \to \eta_i(t),\ i = 1, 2, \ldots, G$ uniformly over the compact set $\mathscr{T}$ so that the classification vector functions $\mathbf{T}_g(x)$, $g = 1, \ldots, G$ will tend to

$$\mathbf{T}_g^*(x) = [T_{g,1}^*(x), \ldots, T_{g,g-1}^*(x), T_{g,g+1}^*(x), \ldots, T_{g,G}^*(x)]^T, \tag{18.4}$$

where $T_{g,i}^*(x) = \mathrm{FCD}^2(x, \eta_i) - \mathrm{FCD}^2(x, \eta_g)$ for $i = 1, \ldots, g-1, g+1, \ldots, G$.

For further discussion, let $\mathscr{C}_i$ denote Class i for $i = 1, \ldots, G$. The prior probabilities of Class i can then be expressed as $\pi_i = \Pr(x \in \mathscr{C}_i)$. For a G-class classification problem, a mistake is made when $x \in \mathscr{C}_g$, by using the classifier, we

assign it to Class i, $i \neq g$. Therefore, the MER of the G-class FCD-based centroid classifier can then be expressed as $\text{MER} = \sum_{g=1}^{G} \pi_g \Big(1 - \Pr\{\mathbf{T}_g(x) > \mathbf{0} | x \in \mathscr{C}_g\}\Big) = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{T}_g(x) > \mathbf{0} | x \in \mathscr{C}_g\}$. The asymptotic MER of the FCD-based centroid classifier is presented in Theorem 18.1.

Theorem 18.1 *Assume the G training functional samples (18.2) are independent with $\text{tr}(\gamma_i) < \infty, i = 1, \ldots, G$. In addition, as $n \to \infty$, we have $n_i/n \to \tau_i > 0$. Then as $n \to \infty$, we have the following asymptotic MER of the FCD-based centroid classifier:*

$$MER \to MER^* = 1 - \sum_{g=1}^{G} \pi_g F_{\mathbf{R}_g}(\boldsymbol{\Sigma}_g^{-1/2} \boldsymbol{\mu}_g), \tag{18.5}$$

where for $g = 1, \ldots, G$, $\boldsymbol{\mu}_g = [\mu_{g,1}, \ldots, \mu_{g,g-1}, \mu_{g,g+1}, \ldots, \mu_{g,G}]^T$, and $\boldsymbol{\Sigma}_g = (\sigma^2_{g_i,g_l}) : (G-1) \times (G-1)$, with $\mu_{g,i} = \|\eta_g\| FCD^2(\eta_i, \eta_g), i = 1, \ldots, g-1, g+1, \ldots, G$, and $\sigma^2_{g_i,g_l} = 4 \int_{\mathscr{T}} \int_{\mathscr{T}} [\tilde{\eta}_i(s) - \tilde{\eta}_g(s)] \gamma_g(s,t) [\tilde{\eta}_l(t) - \tilde{\eta}_g(t)] ds dt, i, l \in \{1, \ldots, g-1, g+1, \ldots, G\}$. In addition, $F_{\mathbf{R}_g}(\cdot), g = 1, \ldots, G$ denotes the cumulative distribution functions of some random variable $\mathbf{R}_g$ which has zero mean vector $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$.

Remark 18.2 The expression (18.5) indicates that the asymptotic MER may not tend to 0 even when the group sample sizes tend to infinity. Note that when MER is 0, there is a perfect classification. However, whether we can have a perfect classification is determined by the data information. If the data are not separable, we cannot have a perfect classification even when the sizes of training samples diverge.

When $G = 2$, the G-class FCD-based centroid classifier reduces to a two-class one. In this case, the results in Theorem 18.1 can be simplified. In addition, we can give an upper error bound of the associated MER. We now denote $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$. The classification function of the two-class FCD-based centroid classifier can then be simply expressed as

$$T(x) = \text{FCD}^2(x, \bar{x}_2) - \text{FCD}^2(x, \bar{x}_1). \tag{18.6}$$

As $n_i, i = 1, 2$ tend to infinity with $n_1/n \to \tau > 0$, $T(x)$ will tend to

$$T^*(x) = \text{FCD}^2(x, \eta_2) - \text{FCD}^2(x, \eta_1). \tag{18.7}$$

Therefore, the MER of the two-class FCD-based centroid classifier $T(x)$ can then be expressed as $\text{MER} = \pi \Pr\{T(x) \leq 0 | x \in \mathscr{C}_1\} + (1 - \pi) \Pr\{T(x) > 0 | x \in \mathscr{C}_2\}$. By Theorem 18.1, we present the asymptotic MER of the two-class FCD-based centroid classifier and its upper bound in Theorem 18.3 below.

Theorem 18.3 *Assume the $(G = 2)$ training functional samples (18.2) are independent with $\text{tr}(\gamma_i) < \infty, i = 1, 2$. In addition, as $n \to \infty$, we have $n_1/n \to \tau > 0$. Then as $n \to \infty$, when we use the FCD-based centroid classifier, we have*

$$MER \to MER^* = \pi F_{R_1}(-\mu_1/\sigma_1) + (1-\pi)[1 - F_{R_2}(\mu_2/\sigma_2)], \tag{18.8}$$

where $\mu_i = \|\eta_i\| FCD^2(\eta_1, \eta_2)$, *and* $\sigma_i^2 = 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_i(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt, i = 1, 2$. $F_{R_i}(\cdot)$ *is the cumulative distribution function of some random variable* R_i *which has mean* 0 *and variance* 1. *Further, the upper bound of the asymptotical MER (18.8) is given by the following expression*

$$MER^* \le \pi F_{R_1}\left(-\frac{\|\eta_1\| FCD(\eta_1, \eta_2)}{2\sqrt{\lambda_{1,\max}}}\right) + (1-\pi)\left[1 - F_{R_2}\left(\frac{\|\eta_2\| FCD(\eta_1, \eta_2)}{2\sqrt{\lambda_{2,\max}}}\right)\right], \tag{18.9}$$

where $\lambda_{i,\max}$ *denote the largest eigenvalue of* $\gamma_i(s,t)$ *for* $i = 1, 2$. *In particular, when the functional data are Gaussian,* $F_{R_1}(\cdot)$ *and* $F_{R_2}(\cdot)$ *should also be replaced with* $\Phi(\cdot)$, *the cumulative distribution function of the standard normal distribution.*

Remark 18.4 The asymptotic MER (18.8) will become smaller if the group mean functions $\eta_1(t)$ and $\eta_2(t)$ become less similar from each other, that is, $FCD(\eta_1, \eta_2)$ becomes larger. This is reasonable. If the group mean functions are not similar, it is easy to classify the new coming observation correctly. In addition, the upper bound of the asymptotic MER (18.9) indicates the smaller the value of $\lambda_{i,\max}, i = 1, 2$ are, the smaller the value of MER^*. This is also reasonable since when $\lambda_{i,\max}, i = 1, 2$ are small, the data are less noisy. Thus, it is easier to classify the new coming functional observation $x(t)$ correctly.

Remark 18.5 If the data are Gaussian, the expression (18.9) indicates that for Gaussian functional data, we always have $MER^* < 1/2$ as long as $FCD(\eta_1, \eta_2) > 0$. That is, the worse case of this two-class FCD-based centroid classifier is better than of the random guessing.

18.5 A Simulation Study

To demonstrate the good performance of the proposed cosine similarity-based classifiers for functional data, we conduct a simulation study in this section. The results of the simulation study allow us to compare the proposed FCD-based centroid and kNN classifiers against some existing centroid and kNN classifiers based on other dissimilarity measures. The centroid and kNN classifiers are defined similarly to the FCD-based centroid and kNN classifiers for functional data as in Sects. 18.4.1 and 18.4.2 except replacing the FCD with one of the dissimilarity measures reviewed in Sect. 18.2. These dissimilarity measures include the L^p-distances for $p = 1, 2$, and ∞, the functional Mahalanobis (FM) semi-distance assuming a common covariance function, and the functional principal components (FPC) semi-distance assuming a common covariance function, as defined in Sect. 18.2. The resulting centroid or kNN classifiers are labeled with L^1, L^2, L^∞, FPC, and FM respectively.

We consider generating functional data for a two-class classification problem under four different scenarios. In the first scenario, two functional samples are generated from two Gaussian processes defined over $I = [0, 1]$, with different group mean functions $\eta_1(t) = 25t^{1.1}(1-t)$ and $\eta_2(t) = 25t(1-t)^{1.1}$ but their covariance functions $\gamma_1(s,t)$ and $\gamma_2(s,t)$ are the same, denoted as $\gamma(s,t)$ whose eigenfunctions are given by $\phi_r(t) = \sqrt{2}\sin(r\pi t)$, $r = 1, 2, \ldots$ and the associated eigenvalues are given by $\lambda_r = 1/(r\pi)^2$, for $r = 1, 2, \ldots$. The generated functions are evaluated at 1000 equidistant time points over $I = [0, 1]$. In the second scenario, the functions are generated in a similar way except that the two covariance functions $\gamma_1(s,t)$ and $\gamma_2(s,t)$ are not the same although their eigenfunctions are the same as those defined in the first scenario but their eigenvalues are given by $\lambda_{1r} = 1/(r\pi)^2$ and $\lambda_{2r} = 2/(r\pi)^2$, for $r = 1, 2, \ldots$ respectively. In the third and fourth scenarios, the functions are generated in a similar way as in the first and second scenarios respectively except the two Gaussian processes are replaced with two standardized exponential processes with rate 1 with the same group mean functions and the group covariance functions.

Under each scenario, two functional samples of equal sizes 100 are generated. The training sample is formed via selecting 50 functions from each sample so that the whole training sample consists of 100 functional observations. The remaining functional observations from the two functional samples form the test sample. The training sample is used to determine the tuning parameters. In particular, we use the 10-fold cross-validation approach. For a kNN classifier, the possible number of nearest neighbors k ranges from 1 to 25. In order to avoid ties, we also set k to be odd numbers only. Similarly, the number of principal components q used in the centroid or kNN classifiers ranges from 1 to q_0 where q_0 may be chosen such that the sum of the first q_0 eigenvalues of the pooled sample covariance function $\widehat{\gamma}(s,t)$ is about 95% of the total variation given by $\mathrm{tr}(\widehat{\gamma})$. Note that the accuracy of a centroid or kNN classifier is measured by its MER which is estimated using the test sample. We repeat the process 1000 times so that we have 1000 MERs. The boxplots of the 1000 MERs of the test samples under the four scenarios are shown in Fig. 18.1.

In view of this figure, it is seen that under the fourth scenario, FCD-based centroid classifier outperforms other centroid classifiers and the FCD-based kNN classifier outperforms other kNN classifiers as well. In the third scenario, the best performance is attended by the proposed FCD-based centroid classifier. Therefore, Gaussianity is not necessarily an advantage for the FCD-based classifiers and they perform well for non-Gaussian data. In practice, it is usually very difficult to check the Gaussianity, hence our proposed classifiers may work well in real problems. In addition, in the first and second scenarios, our FCD-based classifiers perform the second best and FM-based classifiers perform best. However, the FM semi-distance is a rather complicated dissimilarity measure and consumes time in programing and computing.

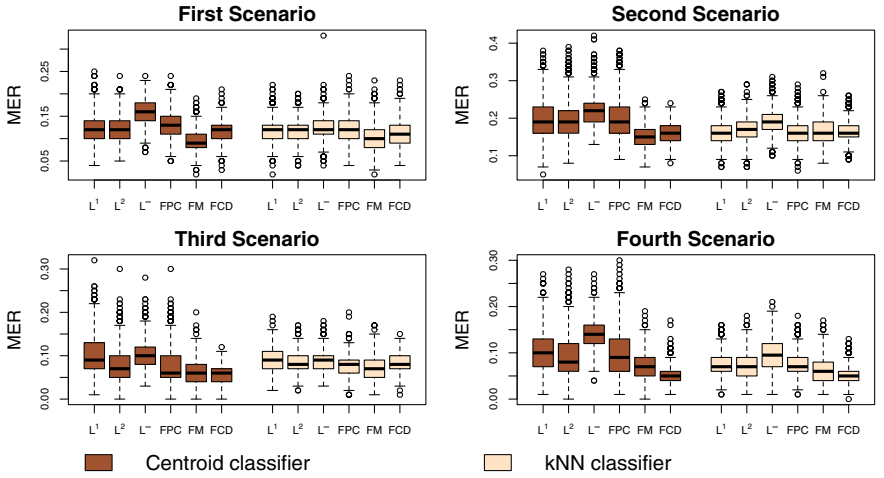

Fig. 18.1 MERs achieved by various centroid and kNN classifiers under all four scenarios

18.6 Application to Australian Rainfall Data

The Australian rainfall data set is available at https://rda.ucar.edu/datasets/ds482.1/. It has been analyzed by [2, 12] respectively to illustrate their classification methodologies. The data set consists of daily rainfall measurements between January 1840 and December 1990, at each of 191 Australian weather stations. The daily rainfall measurements of a station form a rainfall curve. We then have $N = 191$ rainfall curves. Among the 191 weather stations, $N_1 = 43$ of them are located at the northern Australia and the remaining ones are located at the southern Australia. For each station, for simplicity, we just consider the rainfall over a year, i.e., over $t \in [1, 365]$. As in [2], a rainfall curve for a station is obtained via taking the average of the rainfall at each time point $t \in [1, 365]$ over the years which the station had been operating. The resulting raw rainfall curves are then smoothed using a B-spline basis of order 6. The order of B-spline basis is chosen by leave-one-out cross-validation so that the raw rainfall curves can be well represented by the smoothed rainfall curves as shown in Fig. 18.2. From this figure, we can see that some of the weather stations, although geographically located in the north, have a rainfall pattern that is typical of the south. Thus, it is not so easy to distinguish the northern rainfall curves from the southern rainfall curves.

To apply the centroid and kNN classifiers for the Australian rainfall data, we randomly split the 191 rainfall curves into a training sample of size n and a test sample of size $191 - n$ and we take $n = 50$. The number of nearest neighbors is bounded by the smaller sample size of the two classes and the maximum number of eigenfunctions is limited to 20. This process is repeated 1000 times so that we have 1000 MERs for each classifier. Figure 18.3 presents the boxplots of the 1000 MERs of the various centroid and kNN classifiers. It is seen that the FCD-based

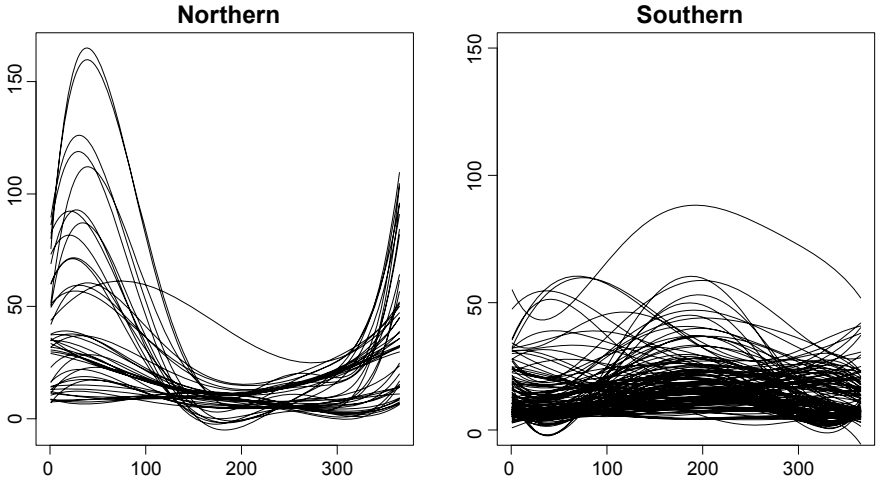

Fig. 18.2 Smoothed Australian rainfall curves for the northern weather stations (left panel) and the southern weather stations (right panel)

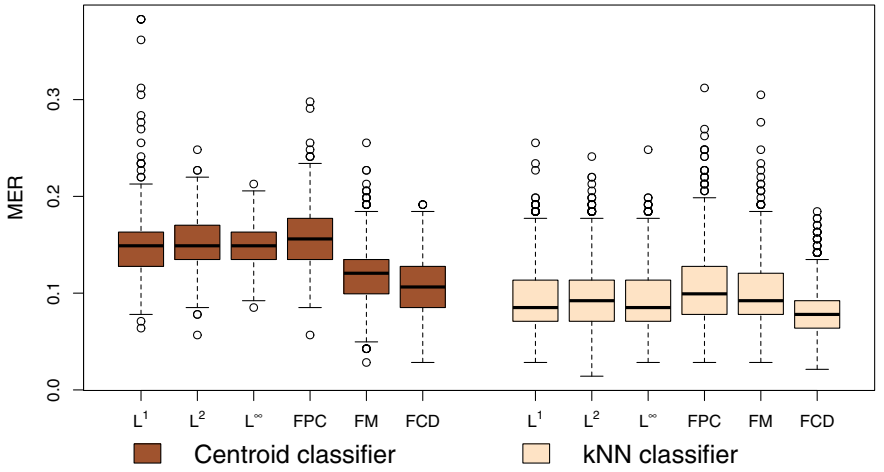

Fig. 18.3 MERs achieved by various centroid and kNN classifiers for the Australian rainfall data

kNN classifier performs best and it obtained mean MER of 0.079. Moreover, the FCD-based centroid classifier outperforms other centroid classifiers which obtained mean MER of 0.106. It is also seen that the kNN classifiers are generally better than the centroid classifiers with the same dissimilarity measures. Using a similar experiment, [3] obtained mean MERs of 0.103 by the centroid classifier which was proposed by [3].

18.7 Concluding Remarks

In this work, we extend the cosine similarity measure for functional data. Based on the FCS, we introduce a new semi-distance for functional data named FCD. This functional semi-distance is simple and can be implemented easily in supervised classification. By combining with the centroid and kNN classifiers, we propose a FCD-based centroid classifier and a FCD-based kNN classifier for functional data. We also study the theoretical properties of the FCD-based centroid classifier. It turns out the cosine similarity-based classifiers for functional data perform well in our simulation study and a real-life data example. As mentioned previously, the range of applications for the new similarity measure or the new functional semi-distance is wide and includes clustering, hypothesis testing, and outlier detection, among others. However, since the proposed FCD does not take the magnitude of the functional data into account, it is expected that the proposed FCD-based classifiers will not perform well for classifying functional data which are different only in their magnitudes. It is then interesting and warranted to study how both the magnitude and shape of the data can be taken into account in FCD-based classifiers so that their performance can be further improved.

18.8 Appendix

Proof (Proof of Theorem 18.1). Under the given conditions, since $\mathrm{tr}(\gamma_i) < \infty$, $i = 1, 2, \ldots, G$, as $n \to \infty$ with $n_i/n \to \tau_i > 0$, we have

$$\mathrm{MER} \to \mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{T}_g^*(x) > \mathbf{0} | x \in \mathscr{C}_g\}, \tag{18.10}$$

where $\mathbf{T}_g^*(x)$ is given in (18.4). Set $\mathbf{S}_g^*(x) = \|x\| \mathbf{T}_g^*(x)$, then we have

$$\mathbf{S}_g^*(x) = [S_{g,1}^*(x), \ldots, S_{g,g-1}^*(x), S_{g,g+1}^*, \ldots, S_{g,G}^*(x)]^T,$$

where $S_{g,i}^*(x) = 2 < x, \tilde{\eta}_g - \tilde{\eta}_i >$, $i = 1, \ldots, g-1, g+1, \ldots, G$. Since $\|x\| > 0$, we have

$$\mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{S}_g^*(x) > \mathbf{0} | x \in \mathscr{C}_g\}. \tag{18.11}$$

When $x \in \mathscr{C}_g$, for $i = 1, \ldots, g-1, g+1, \ldots, G$, we have

$$\mu_{g,i} = \mathbb{E}\left\{S_{g,i}^*(x) | x \in \mathscr{C}_g\right\} = 2 < \eta_g, \tilde{\eta}_g - \tilde{\eta}_i >= \|\eta_g\| \mathrm{FCD}^2(\eta_i, \eta_g),$$

and for any $i, l \in \{1, \ldots, g-1, g+1, \ldots, G\}$,

$$\sigma^2_{g_i,g_l} = \mathrm{Cov}\{S^*_{g,i}(x), S^*_{g,l}(x)|x \in \mathscr{C}_g\}$$
$$= 4\int_{\mathscr{T}}\int_{\mathscr{T}} [\tilde{\eta}_i(s) - \tilde{\eta}_g(s)]\gamma_g(s,t)[\tilde{\eta}_l(t) - \tilde{\eta}_g(t)]dsdt.$$

Therefore, we have

$$\boldsymbol{\mu}_g = \mathbb{E}\{\mathbf{S}^*_g(x)|x \in \mathscr{C}_g\} = [\mu_{g,1}, \ldots, \mu_{g,g-1}, \mu_{g,g+1}, \ldots, \mu_{g,G}]^T,$$
$$\boldsymbol{\Sigma}_g = \mathrm{Cov}\{\mathbf{S}^*_g(x)|x \in \mathscr{C}_g\} = (\sigma^2_{g_i,g_l}) : (G-1) \times (G-1).$$

We can then write $\Pr\{\mathbf{S}^*_g(x) > \mathbf{0}|x \in \mathscr{C}_g\} = \Pr\{\mathbf{R}_g < \boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{\mu}_g|x \in \mathscr{C}_g\}$, where

$$\mathbf{R}_g = \boldsymbol{\Sigma}_g^{-1/2}(-\mathbf{S}^*_g(x) + \boldsymbol{\mu}_g),$$

which is a random variable with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. Therefore,

$$\mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g F_{\mathbf{R}_g}(\boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{\mu}_g), \tag{18.12}$$

as desired where $F_{\mathbf{R}_g}(\cdot)$, $g = 1, \ldots, G$ denote the cumulative distribution functions of $\mathbf{R}_g$, $g = 1, \ldots, G$. □

Proof (Proof of Theorem 18.3) Under Theorem 18.1, when $G = 2$, the classification function of the two-class FCD-based centroid classifier can be simply expressed as (18.6). As $n_i, i = 1, 2$ tend to infinity with $n_1/n \to \tau > 0$, $T(x)$ will tend to (18.7). Thus, the corresponding $S^*(x) = 2 < x, \tilde{\eta}_1 - \tilde{\eta}_2 >$ is a one-dimensional random variable.

When $x \in \mathscr{C}_1$, we have

$$\mu_1 = \mathbb{E}\left\{S^*(x)|x \in \mathscr{C}_1\right\} = 2 < \eta_1, \tilde{\eta}_1 - \tilde{\eta}_2 >= \|\eta_1\|\mathrm{FCD}^2(\eta_1, \eta_2),$$
$$\sigma_1^2 = \mathrm{Var}\left\{S^*(x)|x \in \mathscr{C}_1\right\} = 4\int_{\mathscr{T}}\int_{\mathscr{T}} [\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_1(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt.$$

We can then write $\Pr\{S^*(x) \le 0|x \in \mathscr{C}_1\} = \Pr(R_1 \le -\mu_1/\sigma_1)$ where

$$R_1 = (S^*(x) - \mu_1)/\sigma_1,$$

which is a random variable with mean 0 and variance 1. Similarly, we can show that $\Pr\{S^*(x) > 0|x \in \mathscr{C}_2\} = \Pr(R_2 > \mu_2/\sigma_2)$ where

$$R_2 = (S^*(x) + \mu_2)/\sigma_2,$$

which is a random variable with mean 0 and variance 1, and

$$\begin{aligned}\mu_2 &= -\mathbb{E}\left\{S^*(x)|x \in \mathscr{C}_2\right\}\\ &= -2 < \eta_2, \tilde{\eta}_1 - \tilde{\eta}_2 >= \|\eta_2\|\mathrm{FCD}^2(\eta_1, \eta_2),\\ \sigma_2^2 &= \mathrm{Var}\left\{S^*(x)|x \in \mathscr{C}_2\right\}\\ &= 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_2(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt.\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{MER}^* &= \pi\,\Pr\{S^*(x) \le 0|x \in \mathscr{C}_1\} + (1-\pi)\Pr\{S^*(x) > 0|x \in \mathscr{C}_2\}\\ &= \pi\,\Pr(R_1 \le -\mu_1/\sigma_1) + (1-\pi)\Pr(R_2 > \mu_2/\sigma_2)\\ &= \pi F_{R_1}(-\mu_1/\sigma_1) + (1-\pi)\left[1 - F_{R_2}(\mu_2/\sigma_2)\right],\end{aligned} \tag{18.13}$$

as desired where $F_{R_i}(\cdot), i = 1, 2$ denote the cumulative distribution functions of $R_i, i = 1, 2$.

Let $\lambda_{i,\max}$ denote the largest eigenvalue of $\gamma_i(s,t)$ for $i = 1, 2$. Then we have

$$\begin{aligned}\sigma_i^2 &= 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_i(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt\\ &\le 4\lambda_{i,\max}\|\tilde{\eta}_1 - \tilde{\eta}_2\|^2 = 4\lambda_{i,\max}\mathrm{FCD}^2(\eta_1, \eta_2), \quad i = 1, 2.\end{aligned}$$

It follows that

$$\mu_i/\sigma_i \ge \frac{\|\eta_i\|\mathrm{FCD}^2(\eta_1,\eta_2)}{\sqrt{4\lambda_{i,\max}\mathrm{FCD}^2(\eta_1,\eta_2)}} = \frac{\|\eta_i\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}},\ i = 1, 2.$$

Therefore, by (18.13), we have

$$\mathrm{MER}^* \le \pi F_{R_1}\left(-\frac{\|\eta_1\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}}\right) + (1-\pi)\left[1 - F_{R_2}\left(\frac{\|\eta_2\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}}\right)\right]. \tag{18.14}$$

When the functional data are Gaussian, we have $R_i \sim N(0,1), i = 1, 2$. Therefore, we should replace $F_{R_i}(\cdot), i = 1, 2$ in the expressions (18.13) and (18.14) with $\Phi(\cdot)$, the cumulative distribution of the standard normal distribution. □

References

1. Biau, G., Bunea, F., Wegkamp, M.H.: Functional classification in hilbert spaces. IEEE Trans. Inf. Theory **51**(6), 2163–2172 (2005). https://doi.org/10.1109/TIT.2005.847705
2. Delaigle, A., Hall, P.: Defining probability density for a distribution of random functions. Ann. Stat. **38**(2), 1171–1193 (2010)
3. Delaigle, A., Hall, P.: Achieving near perfect classification for functional data. J. R. Stat. Soc.: Ser. B (Statistical Methodology) **74**(2), 267–286 (2012)

4. Epifanio, I.: Shape descriptors for classification of functional data. Technometrics **50**(3), 284–294 (2008)
5. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics. Springer, New York (2006)
6. Fix, E., Hodges Jr, J.L.: Discriminatory analysis: nonparametric discrimination: consistency properties. US Air Force School of Aviation Medicine. Technical report, vol. 4(3), 477+ (1951)
7. Galeano, P., Joseph, E., Lillo, R.E.: The mahalanobis distance for functional data with applications to classification. Technometrics **57**(2), 281–291 (2015)
8. Glendinning, R.H., Herbert, R.: Shape classification using smooth principal components. Pattern Recogn. Lett. **24**(12), 2021–2030 (2003)
9. Hall, P., Poskitt, D.S., Presnell, B.: A functional datałanalytic approach to signal discrimination. Technometrics **43**(1), 1–9 (2001)
10. Huang, D.S., Zheng, C.H.: Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics **22**(15), 1855–1862 (2006)
11. James, G.M., Hastie, T.J.: Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(3), 533–550 (2001)
12. Lavery, B., Joung, G., Nicholls, N.: A historical rainfall data set for Australia. Australian Meteorol. Mag. **46** (1997)
13. Ramsay, J., Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer, New York (2005)
14. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer Series in Statistics. Springer, New York (2002)
15. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing **69**(7), 730–742 (2006)
16. Sguera, C., Galeano, P., Lillo, R.: Spatial depth-based classification for functional data. Test **23**(4), 725–750 (2014)
17. Song, J.J., Deng, W., Lee, H.J., Kwon, D.: Optimal classification for time-course gene expression data using functional data analysis. Comput. Biol. Chem. **32**(6), 426–432 (2008)
18. Wahba, G.: Spline models for observational data. Soc. Ind. Appl. Math. **59** (1990)